CROSS-RELIABILITY BENCHMARK TEST FOR PRESERVING EMOTIONAL CONTENT IN SPEECH–SYNTHESIS RELATED DATASETS

Jan Hintz[1], *Andreas Wendemuth*[2], *Ingo Siegert*[1]
[1]*Mobile Dialog Systems,* [2]*Cognitive Systems Group,*
IIKT, Otto von Guericke University Magdeburg, Germany
(firstname.lastname)@ovgu.de

Abstract: Emotions play a crucial role in human-machine interaction (HMI), and their accurate representation in speech recordings is essential for creating natural and realistic affective computing components as speech emotion recognition and speech synthesis. However, evaluating the emotional content in speech is a difficult task, as there exist a vast amount of different emotional representations and there is no objective benchmark test to assess the cross-reliability of emotions in different datasets for the HMI domain. This paper evaluates the cross-reliability of emotional content using speech emotion recognition and valence-arousal-dominance prediction models. The study examines three emotional speech datasets, which were selected to represent a range of emotional content as well as different languages (English and German) and are developed in for speech synthesis task. Thereby, the paper especially focuses on the recently published Thorsten emotion dataset.
The results of the conducted experiments showed that the Thorsten emotion dataset achieves state-of-the-art recognition rates on within corpus tests. The experiments also showed high cross-reliability of shared labels (happy/amused, neutral, angry) while unusual labels (drunk, drowsy, whispering) lead to higher confusion.

1 Introduction

Data scarcity is a big problem in data driven domains such as speech emotion recognition (SER) and presents a long known but still urgent challenge [1]. This turns out to be especially true when looking for public available data that should still meet the requirements of high quality recordings, reliable emotion labels, and correct language domain [2]. There are few commonly used speech emotion datasets with recent appearance for German and English: IEMOCAP [3], RAVDESS [4] (English), EmoDB [5] and FAU Aibo Emotion [6] (Currently not accessible) (German). In 2021, the Thorsten emotion dataset [7] with vast amounts of speech recordings from one speaker was released. This is particularly valuable when it comes to emotion-preserving synthesis tasks, as for this purpose high quality recordings of different emotional expressions are needed from one speaker [8].

To this end, reliability has to be ensured in the sense that annotated, or acted emotions are actually perceived, something that has yet not been done for this dataset of elicited emotions. Therefore, this paper performs two recognition experiments, that serve as a benchmark test. An emotion recognition is conducted with data from the same (within-corpus) and from comparable (cross-corpus) datasets. The measurement of the emotion recognition performance using well-known datasets will serve as an objective comparison, but is known to be challenging. Especially when having different emotional categories per dataset, see [9] for a detailed discussion. In this paper, we therefore train and evaluate the performances of emotion recognizers for German and English speech on Thorsten emotion dataset and compare the results in a within-corpus and cross-corpus manner. Additionally, we extract valence, dominance, and arousal labels to gain further insights on the perception of the demonstrated emotions.

2 Related Work

Cross-corpus-reliability is an ongoing research topic in the field of SER. Schuller et al. [10] tested six databases in a cross-corpora evaluation experiment using different types of normalization. The authors observed performance inferiority of cross to within-corpus testing. In [9] a similarity measure is introduced. The authors combined different corpora based on their measured similarity and showed that there is a relation between measured similarity, partly traced back to similar recording conditions, language, and type of emotion, and the recognition performance.

Furthermore, many efforts have been made to adapt SER to different domains [11, 12, 13, 14] However, most of the research that has been conducted on different datasets has focused on comparing datasets across different language domains, and sometimes even comparing datasets containing speakers of different ages. Zhang et al. [15] found that classification accuracy on cross-domain (speech or song) datasets is highest when information is shared between closely related tasks and the output of disparate models are fused. Milner et al. [16] investigated whether information learned from acted emotions is useful for detecting natural emotions, showing the transfer of information is not successful. The authors also highlight the benefits from training on different corpora. They also mention difficulties arising from different methods of annotating emotions, causing a drop in performance. Other approaches [17, 18] found domain adversarial training to be more suitable for generalizing to emotions across datasets. Zhao et al. [19] showed that extracted deep representations combined with a linear support vector classifier are comparable to standard acoustic feature representations in emotion recognition tasks.

3 Method

DATASETS: RAVDESS and EmoDB have been used as benchmark corpora in different experiments related to speech synthesis or emotion recognition [20, 21, 22, 23, 24, 25]. The Thorsten emotion dataset has not yet gained as much attention. All datasets were down-sampled to 16 kHz, as they originally come in different sampling rates.

- **RAVDESS [4]** consists of 1 440 speech files of 4 actors (12 female, 12 male) and contains the eight emotion categories: happy, calm, neutral, sad, angry, fear, surprise, and disgust (●●●●●●●●). The recordings are approx. 4sec long, consisting of two lexically-matched statements, and were recorded at 48kHz sampling rate.
- **EmoDB [5]** consists of 535 speech files of 10 German speakers (5 female and 5 male) and contains the seven emotion categories: happiness, boredom, neutral, sadness, fear, anger, and disgust(●●●●●●●) realized on a neutral speech content. The data was recorded at a 48 kHz sampling rate and then down-sampled to 16 kHz. The recordings are approx. 3sec long.
- **Thorsten emotion dataset (Thorsten) [7]** consists of 2 400 files of a single male German speaker (no actor) and contains the eight expression styles: amused, drowsy, neutral, drunk, whispering, angry, surprised and disgusted(●●●●●●●●). Each emotion has been recorded on 300 identical phrases. The samples range from 2 to 6 seconds. The data was recorded at a 16 kHz sampling rate.

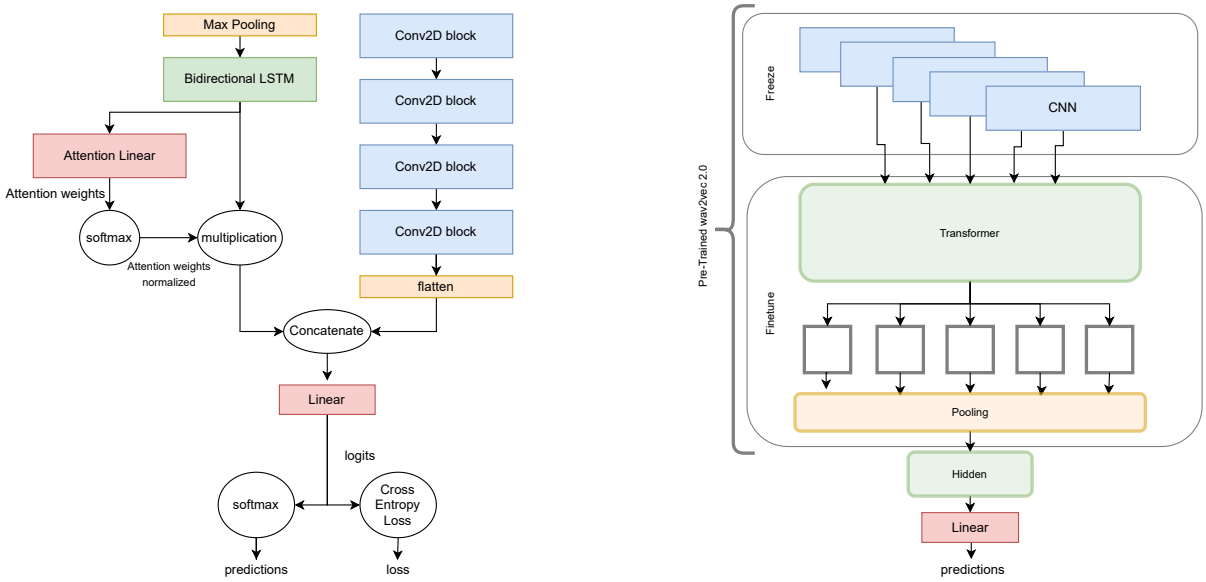
The distribution of the different emotion classes for each of the considered datasets is depicted in Figure 1. It can be seen that RAVDESS and EmoDB share six emotion categories (neutral, happy, disgust, angry, fear, and sad). The category calm (RAVDESS) and bored (EmoDB) are at least similar in their valence and arousal values [26]. All three datasets only share three categories, namely neutral, disgust and angry, while again happy (RAVDESS and EMODB) and amused (Thorsten) can be seen as similar. Additionally, RAVDESS and Thorsten share the category surprise.

(a) RAVDESS Emotion Distribution **(b)** EmoDB Emotion Distribution **(c)** Thorsten emotion Distribution

Figure 1 – Pie chart to illustrate the distribution of samples per emotion class for the considered datasets. Similar emotions are depicted using similar colors.

EVALUATION: To gain more insights and facilitate better understanding of the emotion labels, in addition to a category-based SER also dimensional SER regarding valence, arousal, and dominance were conducted. To objectively determine these dimensions, we used a dimensional SER model (see Figure 2b) introduced by Wagner et al. [27]. This model is based on a pretrained wav2vec 2.0 model [28] (63k hours of English speech). A pooling layer is applied over the hidden states of the last transformer layer, fed through a hidden layer and passed to the final output layer. Wagner et al. fine-tuned the model on the MSP-Podcast corpus. Wagner et al. report a concordance correlation coefficient of 63.8% on MSP-Podcast and 44.8% on IEMOCAP (cross-domain data). We apply this method on each dataset.

In order to effectively assess the emotion categories, a SER model proposed in [29] is used. The author developed a parallel 2D CNN – bidirectional LSTM with attention, based on [19]. The final model, as used in this paper, concatenates the output of the convolutional blocks with the bidirectional LSTM together with an applied attention mechanism as seen in Figure 2a. For each dataset, data splits of 80% training, 10% validation and 10% test were used. During training, the data is augmented with white Gaussian noise. Afterward, models were tested across datasets. The comparison of cross-dataset test is done qualitatively (confusion plots) since the emotions in the datasets don't correspond exactly. To allow a direct objective comparison, additional models were trained on shared labels (happy/amused, angry, neutral, disgust), reporting model accuracy for comparability with state-of-the-art recognition rates.

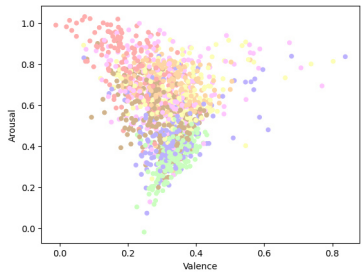


(a) Model for Emotion Prediction (Architecture proposed by Kosta[29])

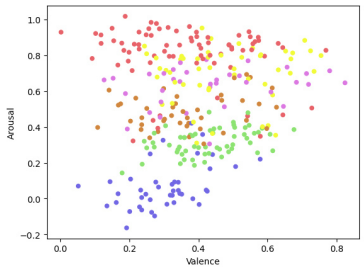
(b) Model for Valence-Arousal-Dominance detection, (Architecture proposed by Wagner et al.[27])

Figure 2 – The two SER model architectures used in this paper.

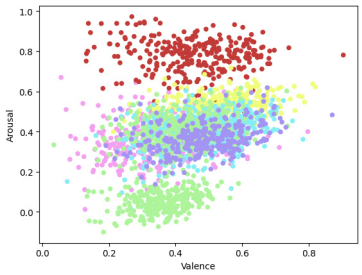
4 Results



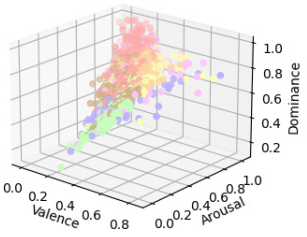
(a) RAVDESS scatter plot of emotions: happy, calm, neutral, sad, angry, fear, surprise and disgust ( )



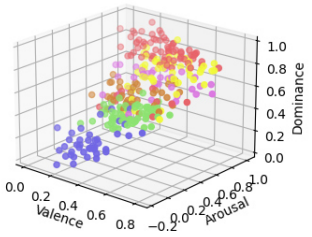
(b) EmoDB scatter plot of emotions: happiness, boredom, neutral, sadness, fear, anger and disgust( )



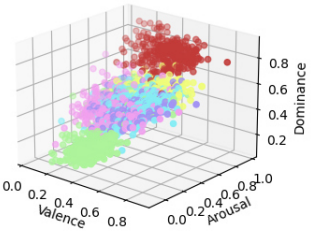
(c) Thorsten scatter plot of emotions: amused, drowsy, neutral, drunk, whispering, angry, surprised and disgusted( ).



(d) RAVDESS 3D Valence-Dominance-Arousal-Plot



(e) EmoDB 3D Valence-Dominance-Arousal-Plot



(f) Thorsten 3D Valence-Dominance-Arousal-Plot

Figure 3 – 2D and 3D Valence-Dominance-Arousal-Plots

The results of the dimensional analysis are shown in Figure 3 as 2D and 3D plots of valence, dominance, and arousal. Among all datasets, the category angry is broadly scattered on the valence axis. It is also the emotion with the highest prediction value for arousal and dominance, followed by happiness and disgust. In case of RAVDESS and Thorsten, calm and sleepy show the lowest arousal values and also the highest density on the valence scale. In EmoDB, this is the case for the category sad. The most clearly defined clusters of emotions can be observed in the 3D-plot of the Thorsten emotion dataset (see. Figure 3f).

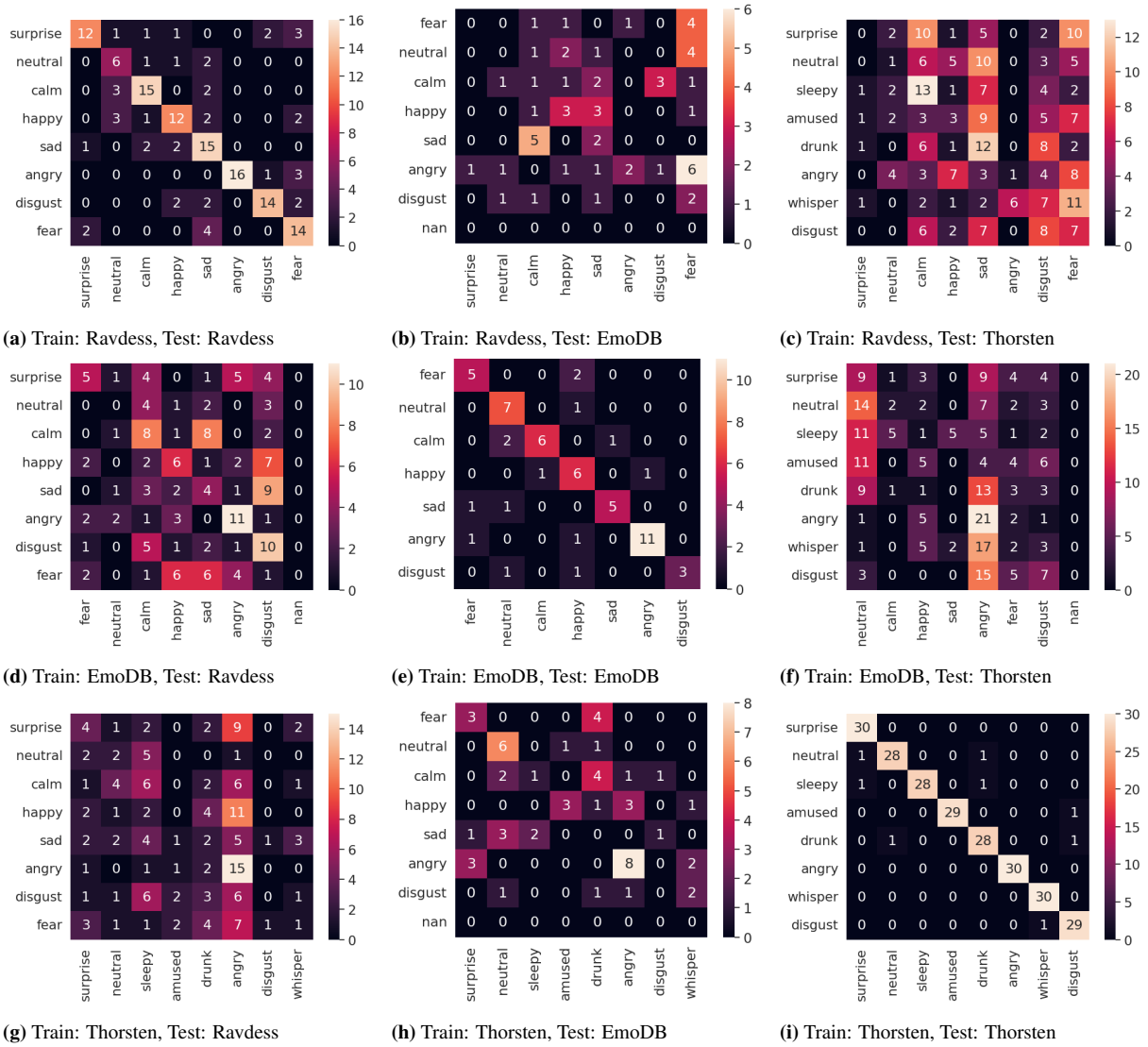
Table 1 – Cross-corpus accuracy on shared labels (happy/amused, angry, neutral, disgust).

Model vs. Test data	RAVDESS	EmoDB	Thorsten
RAVDESS	74.29%	44.12%	33.33%
EmoDB	41.43%	88.24%	44.17%
Thorsten	37.14%	55.88%	97.50%

When tested within corpus, the model trained and tested on Thorsten emotion dataset performed best (96.67% test ACC), followed by the model trained and tested on EmoDB (77.18%) and RAVDESS (71.33% test ACC). On shared labels, within corpus, EmoDB achieves 88.24%, RAVDESS 74.29% and Thorsten emotion Dataset 97.50% test accuracy. The confusion matrices (see. Figure 4) show that the model trained on Thorsten emotion dataset had only minor confusions, while the model trained on EmoDB had some confusion on angry and neutral speech while having a tendency to label samples as neutral or happy. The model trained on RAVDESS had the highest confusion on neutral and surprised, and a high false acceptance on sad.

The results on cross-corpus evaluation on shared labels can be seen in table 1. The confusion matrices also show that not only different labels, but also different language domains, are a cause for increased confusion. In case of the models trained on German datasets (EmoDB and Thorsten), the true positive recognition on English test data (RAVDES) was highest for angry speech (Thorsten 75%, EmoDB 55%). In case of the model trained on the Thorsten emotion dataset, this category also shows the highest false acceptance(25%).

In turn, the model trained on English data has the highest false acceptance rate for fear when tested on EmoDB (31,6%) and Thorsten emotion dataset (21,6%). Across German datasets, the model trained on EmoDB perceived most of Thorsten emotion dataset as angry.



(a) Train: Ravdess, Test: Ravdess **(b)** Train: Ravdess, Test: EmoDB **(c)** Train: Ravdess, Test: Thorsten

(d) Train: EmoDB, Test: Ravdess **(e)** Train: EmoDB, Test: EmoDB **(f)** Train: EmoDB, Test: Thorsten

(g) Train: Thorsten, Test: Ravdess **(h)** Train: Thorsten, Test: EmoDB **(i)** Train: Thorsten, Test: Thorsten

Figure 4 – Confusion plots on Cross-Dataset test data (80% training / 10% validation / 10% test)

5 Conclusion

The common approach of clustering into arousal, valence, and dominance splits confirms that the acted emotions recorded by Thorsten are similarly distributed as in RAVDESS and EmoDB. In comparison to RAVDESS, samples in EmoDB and the Thorsten emotion dataset are more distributed along the valence axis. An effect that can also be seen in the 2D scatter plots dataset. This might be caused by the fact that the utilized model was pretrained and fine-tuned on English

speech data.

It has to be highlighted that in the 3D-scatter plot of the Thorsten emotion dataset, the emotions form clusters that are more coherent than for the other datasets. This is probably because the Thorsten emotion dataset only consists of one speaker. There also seems to be a linear relationship between the dimensions of dominance and arousal, visible in all three datasets, which could hint correlation: lower values of arousal are in accordance with lower values of dominance and vice versa. If this is a true correlation or only caused by accident due to the fact that the training data used for fine-tuning is semi spontaneous, and current the test dataset is acted, has to be analyzed in future experiments.

The Thorsten emotion dataset performs best in terms of both within-corpus testing, reaching state-of-the-art recognition rates (96,67%) and cross-corpus tests on unified labels (97.50%). Unfortunately, the results on RAVDESS reported by [29] (95.40% Acc.) on within-dataset tests could not be reproduced. This might be due to the change of sampling rate, reducing the quality of extracted deep embeddings. The recognition rates on EmoDB of other state-of-the-art models using SVM could not be reproduced as well. This could be due to the fact that EmoDB is a very small dataset of 535 samples vs. Thorsten with 2400 samples and therefore not suitable for a network of this size.

The cross-corpus evaluation also highlights the issue of not having the same emotion labels, as well as different language domains, both leading to high confusion rates. The test with shared labels still showed that the emotion categories contained in the Thorsten emotion dataset achieves similar performance in cross-corpus testing. Especially the additional labels in the Thorsten emotion dataset (i.e., drunk, whispering, sleepy) lead to confusion. Nonetheless, this specific data could prove useful in special scenarios such as driver evaluation, for improving the conditions for safe driving as discussed by [30]. Overall, Thorsten emotion dataset provides, a relatively big amount of high quality speech samples, thus allowing good results in terms of cross reliability.

Acknowledgements

This research has been funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment, and Prevention against Child Sexual Abuse).

References

[1] NIEBUHR, O. and A. MICHAUD: *Speech data acquisition -: The underestimated challenge. Kieler Arbeiten in Linguistik und Phonetik (KALIPHO)*, 3, pp. 1–42, 2015.

[2] SIEGERT, BÖCK, and WENDEMUTH: *Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. Journal on Multimodal User Interfaces*, 8(1), pp. 17–28, 2014. doi:10.1007/s12193-013-0129-9. URL `http://dx.doi.org/10.1007/s12193-013-0129-9`.

[3] BUSSO, C., M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J. N. CHANG, S. LEE, and S. S. NARAYANAN: *Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation*, 42(4), pp. 335–359, 2008.

[4] LIVINGSTONE, S. R. and F. A. RUSSO: *The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one*, 13(5), p. e0196391, 2018.

[5] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. F. SENDLMEIER, B. WEISS ET AL.: *A database of german emotional speech.* In *Interspeech*, vol. 5, pp. 1517–1520. 2005.

[6] BATLINER, A., S. STEIDL, and E. NÖTH: *Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus*. 2008.

[7] MÜLLER, T. and D. KREUTZ: *Thorsten - open german voice (emotional) dataset*. 2021. doi:10.5281/zenodo.5525023. URL `https://doi.org/10.5281/zenodo.5525023`. Please use it to make the world a better place for whole humankind.

[8] ADIGWE, A., N. TITS, K. E. HADDAD, S. OSTADABBAS, and T. DUTOIT: *The emotional voices database: Towards controlling the emotion dimension in voice generation systems*. *arXiv preprint arXiv:1806.09514*, 2018.

[9] SIEGERT, BÖCK, and WENDEMUTH: *Using a pca-based dataset similarity measure to improve cross-corpus emotion recognition*. *Computer Speech & Language*, 51, pp. 1 – 23, 2018. doi:https://doi.org/10.1016/j.csl.2018.02.002. URL `http://www.sciencedirect.com/science/article/pii/S0885230816302650`.

[10] SCHULLER, B., B. VLASENKO, F. EYBEN, M. WÖLLMER, A. STUHLSATZ, A. WENDEMUTH, and G. RIGOLL: *Cross-corpus acoustic emotion recognition: Variances and strategies*. *IEEE Transactions on Affective Computing*, 1(2), pp. 119–131, 2010.

[11] DENG, J., X. XU, Z. ZHANG, S. FRÜHHOLZ, and B. SCHULLER: *Universum autoencoder-based domain adaptation for speech emotion recognition*. *IEEE Signal Processing Letters*, 24(4), pp. 500–504, 2017.

[12] HASSAN, A., R. DAMPER, and M. NIRANJAN: *On acoustic emotion recognition: compensating for covariate shift*. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7), pp. 1458–1468, 2013.

[13] ZONG, Y., W. ZHENG, T. ZHANG, and X. HUANG: *Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression*. *IEEE signal processing letters*, 23(5), pp. 585–589, 2016.

[14] DENG, J., Z. ZHANG, E. MARCHI, and B. SCHULLER: *Sparse autoencoder-based feature transfer learning for speech emotion recognition*. In *2013 humaine association conference on affective computing and intelligent interaction*, pp. 511–516. IEEE, 2013.

[15] ZHANG, B., E. M. PROVOST, and G. ESSL: *Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach*. In *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 5805–5809. IEEE, 2016.

[16] MILNER, R., M. A. JALAL, R. W. NG, and T. HAIN: *A cross-corpus study on speech emotion recognition*. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 304–311. IEEE, 2019.

[17] GIDEON, J., M. G. MCINNIS, and E. M. PROVOST: *Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)*. *IEEE Transactions on Affective Computing*, 12(4), pp. 1055–1068, 2019.

[18] ABDELWAHAB, M. and C. BUSSO: *Domain adversarial for acoustic emotion recognition*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12), pp. 2423–2435, 2018.

[19] ZHAO, Z., Z. BAO, Y. ZHAO, Z. ZHANG, N. CUMMINS, Z. REN, and B. SCHULLER: *Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition*. *IEEE Access*, 7, pp. 97515–97525, 2019.

[20] STEIDL, S., T. POLZEHL, H. T. BUNNELL, Y. DOU, P. K. MUTHUKUMAR, D. PERRY, K. PRAHALLAD, C. VAUGHN, A. W. BLACK, and F. METZE: *Emotion identification for evaluation of synthesized emotional speech*. In *Proc. Speech Prosody 2012*, pp. 661–664. 2012.

[21] HÖBEL-MÜLLER, J., I. SIEGERT, R. HEINEMANN, A. F. REQUARDT, M. TORNOW, and A. WENDEMUTH: *Analysis of the influence of different room acoustics on acoustic emotion features and emotion recognition performance*. In *Tagungsband - DAGA 2019*, pp. 886–889. Rostock, Germany, 2019.

[22] KIENAST, M. and W. F. SENDLMEIER: *Acoustical analysis of spectral and temporal changes in emotional speech*. In *Proc. ITRW on Speech and Emotion*, pp. 92–97. 2000.

[23] XU, M., F. ZHANG, and W. ZHANG: *Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset*. *IEEE Access*, 9, pp. 74539–74549, 2021.

[24] KWON, S. ET AL.: *Att-net: Enhanced emotion recognition system using lightweight self-attention module*. *Applied Soft Computing*, 102, p. 107101, 2021.

[25] PEPINO, L., P. RIERA, and L. FERRER: *Emotion recognition from speech using wav2vec 2.0 embeddings*. *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[26] SCHULLER, B., B. VLASENKO, F. EYBEN, G. RIGOLL, and A. WENDEMUTH: *Acoustic Emotion Recognition: A Benchmark Comparison of Performances*. In *Proc. of the IEEE ASRU-2009*, pp. 552–557. 2009.

[27] WAGNER, J., A. TRIANTAFYLLOPOULOS, H. WIERSTORF, M. SCHMITT, F. BURKHARDT, F. EYBEN, and B. W. SCHULLER: *Dawn of the transformer era in speech emotion recognition: closing the valence gap*. *arXiv preprint arXiv:2203.07378*, 2022.

[28] HSU, W.-N., A. SRIRAM, A. BAEVSKI, T. LIKHOMANENKO, Q. XU, V. PRATAP, J. KAHN, A. LEE, R. COLLOBERT, G. SYNNAEVE ET AL.: *Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training*. *arXiv preprint arXiv:2104.01027*, 2021.

[29] JOVANOVIC, K.: *Speech-emotion-classification-with-pytorch*. `https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch`, 2022.

[30] GRIMM, M., K. KROSCHEL, H. HARRIS, C. NASS, B. SCHULLER, G. RIGOLL, and T. MOOSMAYR: *On the necessity and feasibility of detecting a driver's emotional state while driving*. In A. C. R. PAIVA, R. PRADA, and R. W. PICARD (eds.), *Affective Computing and Intelligent Interaction*, pp. 126–138. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.